New Data Sources:
Accessibility and Use

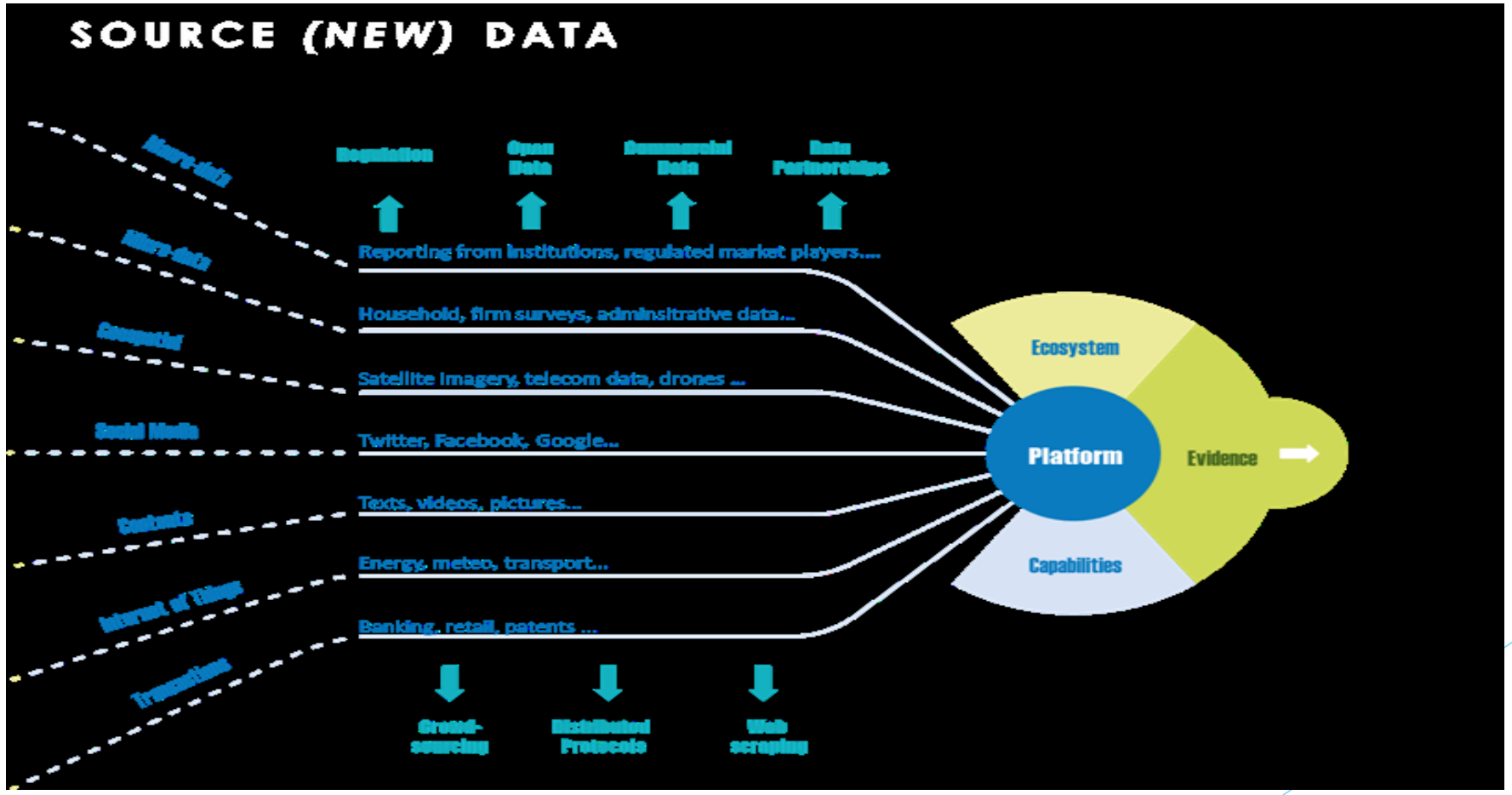
UNECE CES Seminar

Organized by United States and Switzerland

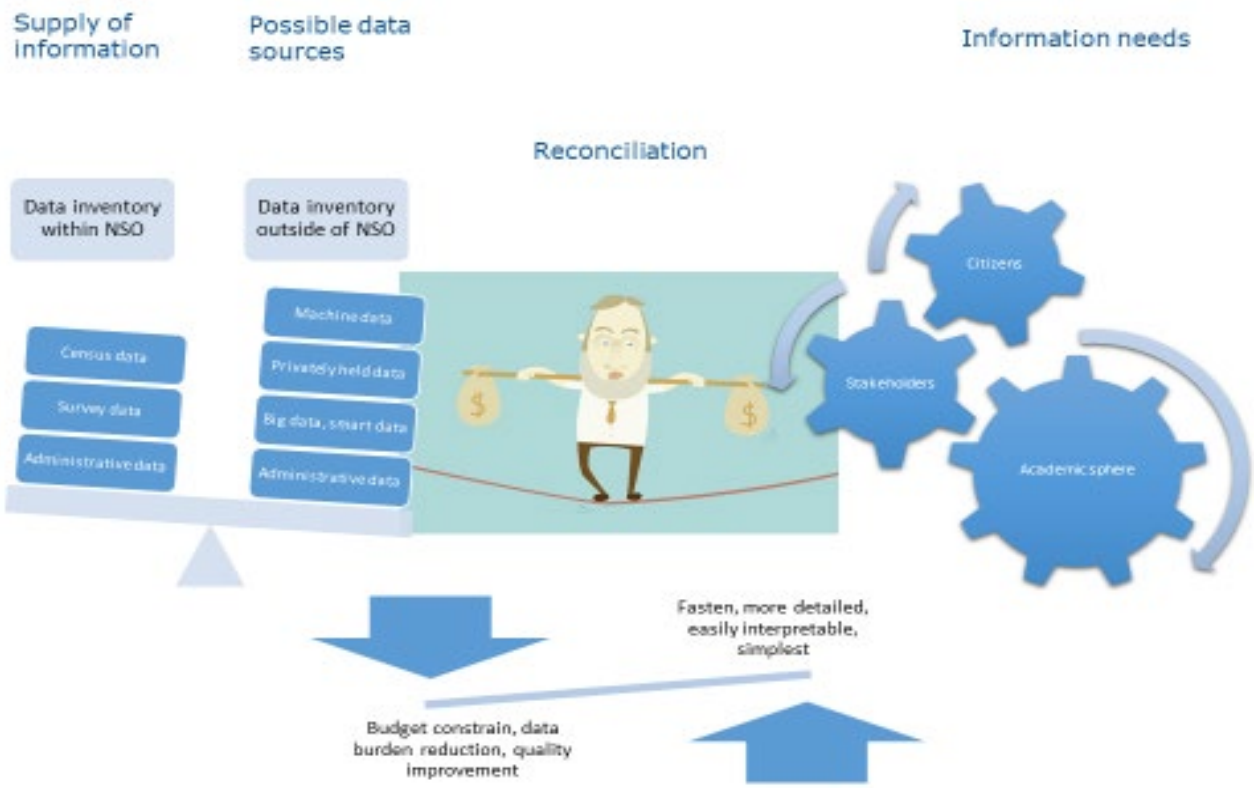
June 2019

Overview of today's seminar

New data sources are everywhere



Organizing ourselves for new data



New skills needed

New data addressing old issues

Plan for today's seminar

- Keynote presentation
- Organized in two sessions – accessing new data sources and skills needed to use new data sources
- Session 1 – accessing new data sources -- led by Switzerland
- Session 2 – skills needed to use new data sources -- led by United States
- After lunch -- sum up, focusing on actionable outcomes

Keynote speaker: Frauke Kreuter

Session 1: Accessing new data sources

Introduction by the Session Chair (Switzerland)

- **Livio Lugano**
  Vice-Director SFSO, Head of Economic Division

- **Dr. Bertrand Loison**
  Vice-Director SFSO, Head of Register Division

Aim of the Session

- The session will consider success stories and challenges faced by national statistical organizations in their efforts to obtain data from alternative sources.

- The discussion will focus on experiences and lessons learned, legal hurdles related to accessing alternative data and forging partnerships to access alternative data.

Contributions

- 7 papers from the NSOs from Germany, Mexico, Norway, Korea, Hungary, Netherlands, Sweden.
- 4 papers from International Organizations Eurostat, UNCTAD, OECD, UNSD.
- Thank you to the contributors who submitted their abstracts and papers in accordance with the requested milestones !

Legal hurdles: Expectations for the session

- The legal hurdles link with the topic « Accessing New Data Sources » is discussed since many years without a clear action plan to solve the problem.
- Accessing New Data Sources » is often a national and/or supranational problem that doesn't match anymore with the legal bases of national scope. It makes the problem complex to solve.

Can we really initiate a concrete actionplan at a UNECE level that can help all NSOs regarding the legal hurdles or should this topic be solved at the national level ? And if yes, how ?

Partnerships: Expectations for the session

- Establishing « Partnerships » to ensure having access to new data sources is a promising but also a painful way for small NSOs.
- Contracts and agreements are often time limited, subject to modifications and need to be usually renegotiated on a regular basis, which can be of real concern to ensure a stable statistics production over the time.

Can we concretely act at the international level to facilitate the establishment of partnerships with international data providers ? And if yes, how ?

Summary of issues raised in the contributed papers

Main goals

Data Eco-system	New statistics & Architecture	Legal Framework
Developing their « own » data eco-system with strong emphasis on establishing public and private partnerships at the national and international levels.	Connecting new data sources to create new statistics and/or supplementing existing statistics. Architecture for Trusted Smart Statistics.	Adapting the national legal framework to ensure having access to all administrative records as well as data from the private sectors.
Ex: Sweden, Korea, …	Ex: Netherlands, Eurostat, Hungary, Germany, UNSD, …	Ex: Korea, Sweden, …

Building a data eco-system and new legal framework seems to be goals shared by the countries.

Pushing new architectures and platforms is the focus of international organizations.

The Second Seminar at the Conference – since it will discuss the Role of NSOs in the new Data ecosystem

Important steps

Competence Center	Assessing the NSO	Improving NSO's Frameworks
Creating a new institution (Data Science or Big Data Campus), dedicated to cope with all issues linked with new data sources.	Evaluating the current situation e.g. Information, Standards & Frameworks, Processes, Institutional arrangements, People, Methods and Systems.	Improving existing NSO's frameworks to integrate the new data sources and methods into the current production and organization.
Ex: Netherlands, UNSD, (UK), Norway, …	Ex: Sweden, Germany, …	Ex: Germany, Hungary, …

The initiated steps to achieve the goals are very diverse. It's a logical consequence of goal's variety that have been set. No real common roadmap shared by all NSO's.

Requirements/Consequences for the NSOs

- A DIFFERENT MINDSET IS NEEDED: a systematic approach, define a strategy, first think than act, set priorities, from questions first to a content analytical approach, alternative ways of producing statistics, be open for results;
- COMMUNICATE more with stakeholders from politics, policy makers, business, academia and NGOs, and international organizations also internal, supported and accepted by the rest of the organization;
- MONITORING: Setup concrete targets for desirable quality in registers, cost-benefit analysis needs to be done, before big data is used for the production of statistics and not only for feasibility studies;
- INITIAL INVESTMENTS: for hard- and software as well as the training of statistical employees for using big data or interfaces for the integration of semi-final products;
- IN CASE OF COOPERATION: the principle of neutrality of official statistics has to be assured and to be secured through legally binding, transparent procedures.

Ex: Netherlands, Korea, Sweden, Norway, Germany, …

How to get access to data? (I)

Cooperation

Willingness to cooperate and to establish partnerships with the private sector 1) to have access to data, 2) to share metadata, 3) to define technical standards, 4) to ensure data privacy, etc.

Ex: Netherlands, Korea, Sweden, Norway, UNSD, Eurostat, …

Obligation

Redefining the legal framework to give the NSO the right to have access to the data of the private sector (people and business). National laws are limited to the enterprises located in the own country.

Ex: (UK), (F), …

Cooperation in form of memorandum of understanding seems to be the dominant model used by the countries to have access to the data from the private sector.

How to get access to data? (II)

- Access to personal information in private sector is limited – strong privacy protection and sector-specific laws, only for the production of official statistics not for studies in big data projects.
- Lack of cooperation from private sector data providers, passive approach on data sharing. Low quality of private sector data, representativity, consistency and completeness.
- Shortage of experts such as data scientists and IT infrastructures. Method of de-identification of personal information.
- Restriction on budgets and inflexible recruitment process.

What are the results ?

Pilot projects

Many projects with different partners from the private and public sector and new data sources.

Ex: Netherlands, Korea, Sweden, Norway, UNSD, UNCTAD, Eurostat, …

Data

More relevant and timely data for decision making through linking various data. Reduce the statistical production cost.

Ex: Korea, …

Infrastructure

Sharing data infrastructure, linkage between public & private sector big data and de-identification services.

Ex: Netherlands, Korea, Sweden, Norway, UNSD, Eurostat, …

Training, Skills, Capacity - building

Several concrete results have been achieved in the area of skills required to process new data sources. This item will be dealt with in the 2nd session.

Questions for authors and delegates

1. **Legal hurdles :** Do we need to initiate a concrete action plan at a UNECE level to ensure that NSO's can have access to data from national interest that are stored by the private sector in a unclear geographical location (cloud)? If yes who should be in charge to define the content of this action plan?

2. **Partnerships:** An important criteria of statistical information include that they are permanently available and comparable internationally. In order for the data to remain internationally comparable, it is important that we use a uniform approach, the same definitions and the same technical standards (Edge Analytics). What are your experiences and how can we ensure this and who should be in charge to establish these?

3. **Finance:** In addition to reducing the burden on respondents by using new data sources, it is also budget restrictions that make us look for new solutions. The initial investments of this new “strategy” are not insignificant. How has this challenge been solved by the various NSOs?

Session 2: Skills needed to use new data sources

- Dr. William Beach, Commissioner, U.S. Bureau of Labor Statistics

Papers contributed

- Two papers focused specifically on this session:
  - Slovenia – New competencies needed to work with data – lessons learnt so far
  - Sweden – Improving data integration with the help of the Global Statistical Geospatial Framework
- An additional paper, while focused on session 1, had a specific section on skills:
  - United Nations Statistics Division – New data sources for official statistics – Access, use and new skills

But really, everyone contributed

- Eurostat
- Germany
- Hungary
- Korea
- Mexico
- Netherlands
- Norway
- OECD
- UNCTAD

Organizing the skills

- How to acquire new data
- Infrastructure to support new data
- How to work with new data

Summary of issues raised in the contributed papers

How to acquire new data

- Identify sources, and the right sources
- Data collection modalities
- Partnerships and other relationships
- Legal environment

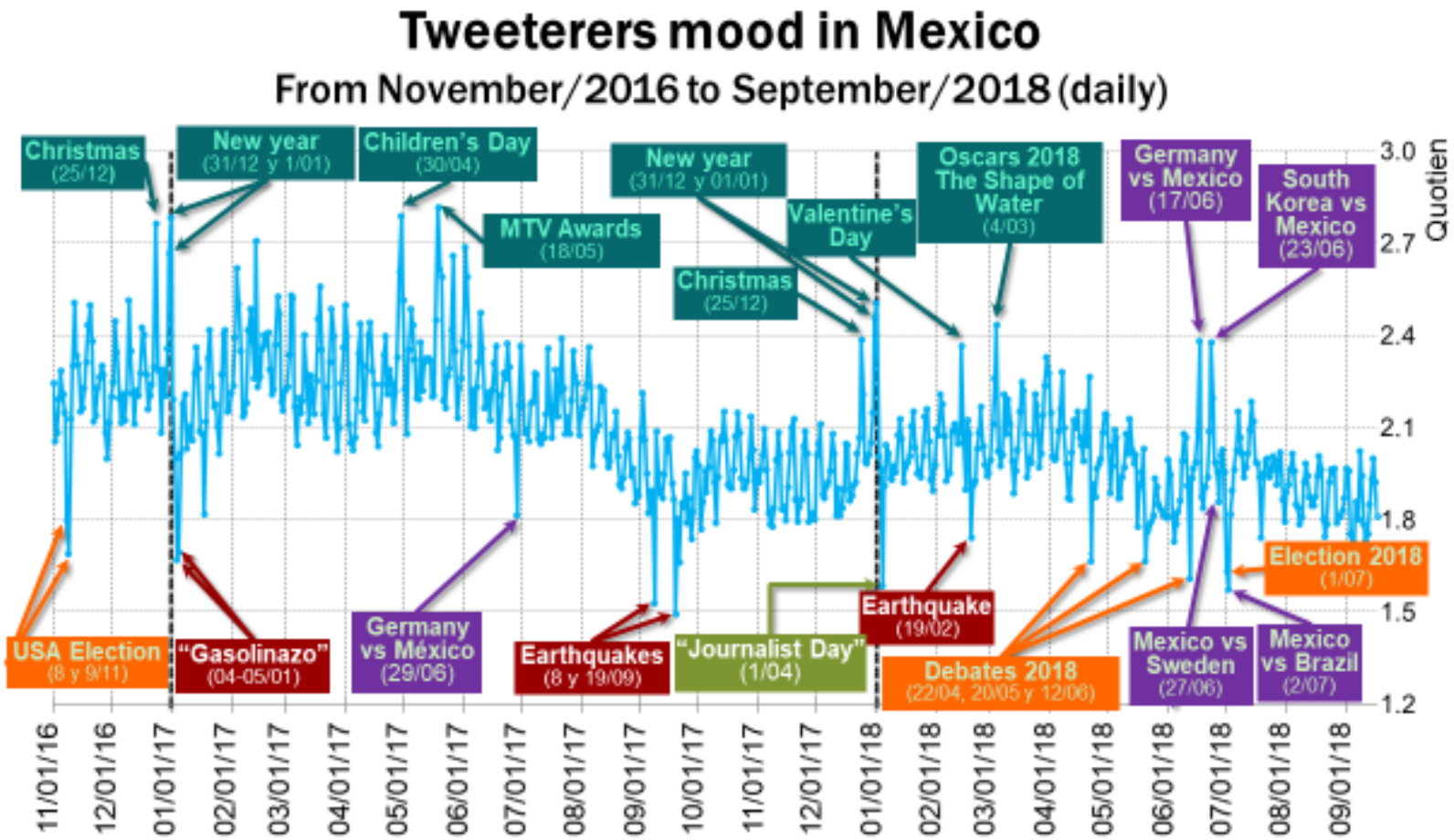
Infrastructure to support new data

- New technologies: .net, asp.net, ml.net frameworks
- Tools – MS visual studio, SQL
- Languages – C#, R, Python
- Databases – MS SQL Server, Oracle
- Python libraries – TensorFlow, Scikit-Learn, Keras

How to work with new data

- Text mining
- Machine learning
- Data modeling
- Natural language processing
- Sentiment analysis techniques

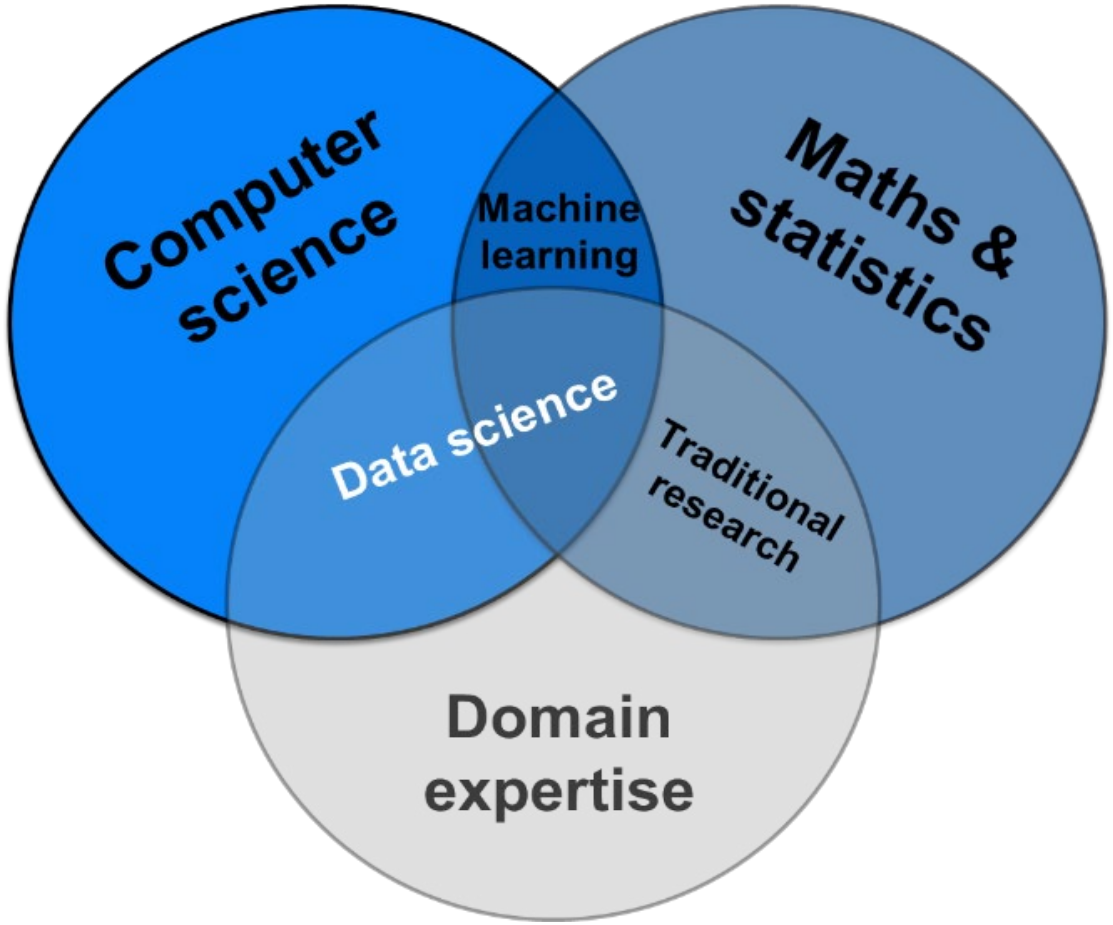
Sentiment analysis of Mexican tweets



How to learn

- Learn by doing
- Participate in international activities
- Self-study; online
- Formal classes

Who will lead the way?



Questions for authors and delegates

Question -- Slovenia

- Your paper mentions many ways that your staff acquired the needed new skills, such as self-study, learning by doing, attending conferences, and more formal studies. You make the strong point that you can't separate training and development from ongoing work, and in fact you want to connect training to specific projects. Can you describe an example where you brought together staff from different areas for a project, the training they received, and the outcome?

Question -- Sweden

- Your paper focuses on the integration of geospatial information into statistical production. You describe a capability assessment to determine if you have the resources available for the work at hand, including people, systems, processes, and other factors. Can you say a little more about how you assess the capability of people for these new data tasks, and what you do when the needed skills are not available?

Question – United Nations Statistics Division

- You wrote about accessing and using very large datasets. You mentioned that one of the lessons-learned is you don't try to transfer the datasets, but rather "perform processing at the place where the data are generated." In fact, you describe a change in paradigm, from "moving the data to the analytics" to "moving the analytics and the data quality framework to the data." Can you provide an example and identify any challenges with this approach? For example, are there concerns about data confidentiality?

Questions for general discussion

- How has the process of obtaining voluntary cooperation changed as you transition from obtaining information directly (in person, from countries) to obtaining information from a website (via APIs or web-scraping software)?
- Business websites often have restrictive terms of service. What experience have you had in working with businesses to obtain data through their website or API and still conform to their terms of service?
- Are you hiring data scientists? What educational background do they have? What skills are you looking for?
- What steps are you taking to get more experienced staff trained in some of the new data techniques?

Key outcomes from Session 1

Outcome 1: Legal hurdles

- A global « Data clearance » should be developed for official statisticians in order to ensure that NSOs will have access to all necessary data from national interest – without consideration of geographical locations or data sensitivity – needed for production.

- This « Data clearance » should also allow NSOs to put and run algorithms into enterprises that store data. This is necessary to move into the Edge Analytics (putting computation out).

Outcome 2: Partnerships

- New data sources (esp. Big Data) do not offer a high security regarding their « permanently availableness ». Therefore the statistical community should develop a methodological handbook that can cope with this issue.

- The statistical community has to be much more proactive regarding the establishment of norms and to take an active role in the definition of international standards (ISO,…).

Key outcomes from Session 2

Outcome 1: Define the skill sets needed to use new data sources in the future

- Countries and organizations should share information on their efforts to identify the jobs and skills needed to use new data sources in the future
  - What is a data scientist?
  - Is it one occupation or many?
  - Identify the variety of skills and disciplines needed
  - Create a curriculum for learning those skills
  - Ensure that current staff aren't left behind
  - Encourage future generations to acquire "data science" skills

Outcome 2: Establish pilot projects that use new data sources

- Define projects that multiple countries and organizations can undertake
  - Share plans
  - Develop timelines for training staff and running pilots
  - Share results
  - Identify lessons learned, best practices

Comments and discussion on key outcomes

Closing comments/conclusions

- New data sources come with a host of issues
  - Discussion of Data Access and Skills overlapped – many common issues
- Areas for information-sharing and further investigation:
  - Legal hurdles
    - Changing laws to allow access
    - Working with businesses to allow access
  - Partnerships
    - Public-private
    - Across countries
  - Training needs
    - How to obtain data (from countries, businesses, individuals)
    - How to work with new data sources

Thank You

Thanks to our keynote speaker, all those who prepared papers, and all of you who joined in a great discussion